A Highly Biased View of the 'ART' of Data Assimilation

Jeffrey L. Anderson

20 March, 2002

I. Overview and some methods

Big problems require clever simplification

II. Challenges

A. Model bias

B. Balances and attractors

C. Assimilation and discrete distributions

III. Opportunities

Field is maturing; theory and methods that are easy to apply

Software engineering advances make it easier to get started

Efforts like Data Assimilation Research Testbed (DART) underway

A. A plethora of untouched models and observations

B. Improved assimilation methods for existing problems

C. Improved use of existing observations; quality control

D. Using data to improve models

E. Evaluating value of existing observations

F. Evaluating future observing systems

H. Adaptive observations

The Data Assimilation Problem

Given:

---

1. A physical system (atmosphere, ocean...)

---

2. Observations of the physical system

   Usually sparse and irregular in time and space
   Instruments have error of which we have a (poor) estimate
   Observations may be of ‘non-state’ quantities
   Many observations may have very low information content

---

3. A model of the physical system

   Usually thought of as approximating time evolution
   Could also be just a model of balance (attractor) relations
   Truncated representation of ‘continuous’ physical system
   Often quasi-regular discretization in space and/or time
   Generally characterized by ‘large’ systematic errors
   May be ergodic with some sort of ‘attractor’

The Data Assimilation Problem (cont.)

We want to increase our information about all three pieces:

---

1. Get an improved estimate of state of physical system

   Includes time evolution and ‘balances’
   Initial conditions for forecasts
   High quality analyses (re-analyses)

---

2. Get better estimates of observing system error characteristics

   Estimate value of existing observations
   Design observing systems that provide increased information

---

3. Improve model of physical system

   Evaluate model systematic errors
   Select appropriate values for model parameters
   Evaluate relative characteristics of different models

Examples:

A. Numerical Weather Prediction
Model: Global troposphere / stratosphere O(1 degree by 50 levels)
Observations: radiosondes twice daily, surface observations, satellite winds, aircraft reports, etc.

B. Tropical Upper Ocean State Estimation (ENSO prediction)
Model: Global (or Pacific Basin) Ocean O(1 degree by 50 levels)
Observations: Surface winds (possibly from atmospheric assimilation), TAO buoys, XBTs, satellite sea surface altimetry

C. Mesoscale simulation and prediction
Model: Regional mesoscale model (WRF), O(1km resolution)
Observations: Radial velocity from Doppler radar returns

D. Global Carbon Sources and Sinks

Nonlinear Filtering

Dynamical system governed by (stochastic) DE

$$dx_t = f(x_t, t) + G(x_t, t)d\beta_t, \qquad t \geq 0 \tag{1}$$

Observations at discrete times

$$y_k = h(x_k, t_k) + v_k; \;\; k = 1, 2, \ldots; \qquad t_{k+1} > t_k \geq t_0 \tag{2}$$

Observational error is white in time and Gaussian

$$v_k \rightarrow N(0, R_k) \tag{3}$$

Complete history of observations is

$$Y_\tau = \{y_l; \;\; t_l \leq \tau\} \tag{4}$$

Goal: Find probability distribution for state at time t

$$p(x, t | Y_t) \tag{5}$$

Nonlinear Filtering (cont.)

State between observation times obtained from DE

Need to update state given new observation

$$p\left(x, t_k \middle| Y_{t_k}\right) = p\left(x, t_k \middle| y_k, Y_{t_{k-1}}\right) \tag{6}$$

Apply Bayes' rule

$$p\left(x, t_k \middle| Y_{t_k}\right) = \frac{p\left(y_k \middle| x_k, Y_{t_{k-1}}\right) p\left(x, t_k \middle| Y_{t_{k-1}}\right)}{p\left(y_k \middle| Y_{t_{k-1}}\right)} \tag{7}$$

Noise is white in time (3) so

$$p\left(y_k \middle| x_k, Y_{t_{k-1}}\right) = p(y_k \middle| x_k) \tag{8}$$

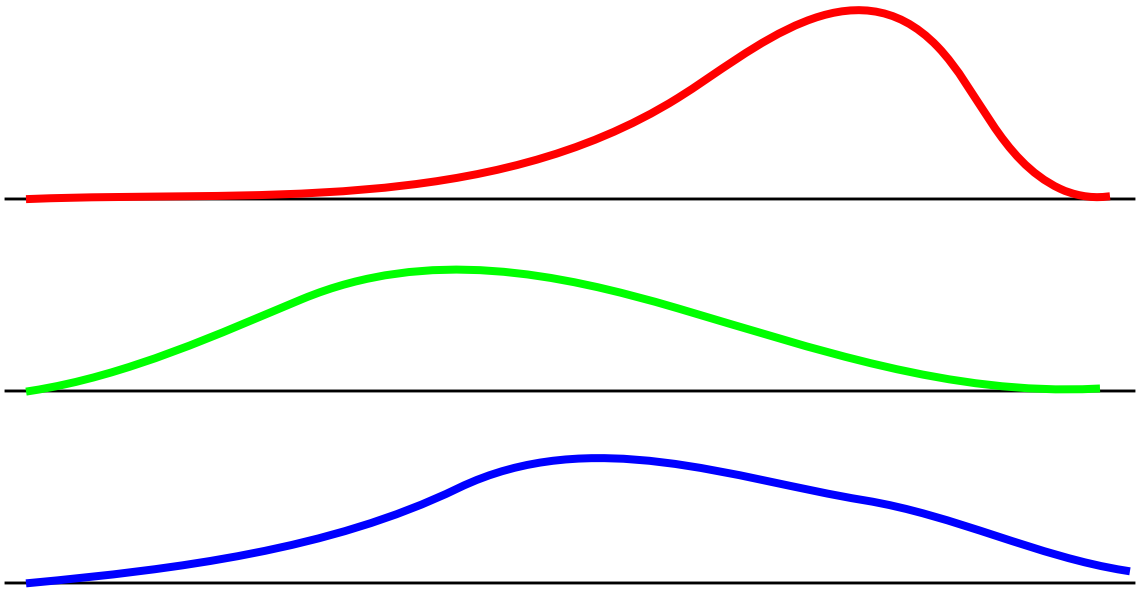
Also have

$$p\left(y_k \middle| Y_{t_{k-1}}\right) = \int p(y_k \middle| x) p\left(x, t_k \middle| Y_{t_{k-1}}\right) dx \tag{9}$$

Nonlinear Filtering (cont.)

Probability after new observation

$$p\left(x, t_k \middle| Y_{t_k}\right) = \frac{p(y_k|x)\, p\left(x, t_k \middle| Y_{t_{k-1}}\right)}{\int p(y_k|\xi)\, p\left(\xi, t_k \middle| Y_{t_{k-1}}\right) d\xi} \qquad (10)$$



Second term in numerator, denominator comes from DE
First term comes from distribution of observational error

General methods for solving the filter equations are known:

1. Advancing state estimate in time

2. Taking product of two distributions

But, these methods are far too expensive for problems of interest

1. Huge model state spaces (10 is big!), NWP models at O(10 million)

2. Need truncated representations of probabilistic state to avoid exponential solution time and storage

The ART of Data Assimilation:

Find heuristic simplifications that make approximate solution affordable

1. Localization (spatial or other truncated basis)

2. Linearization of models, represent time evolution as linear (around a control non-linear trajectory)

3. Represent distributions as Gaussian (or sum of Gaussians)

4. Monte Carlo methods

5. Application of simple balance relations

Kalman Filter

Simplifications:

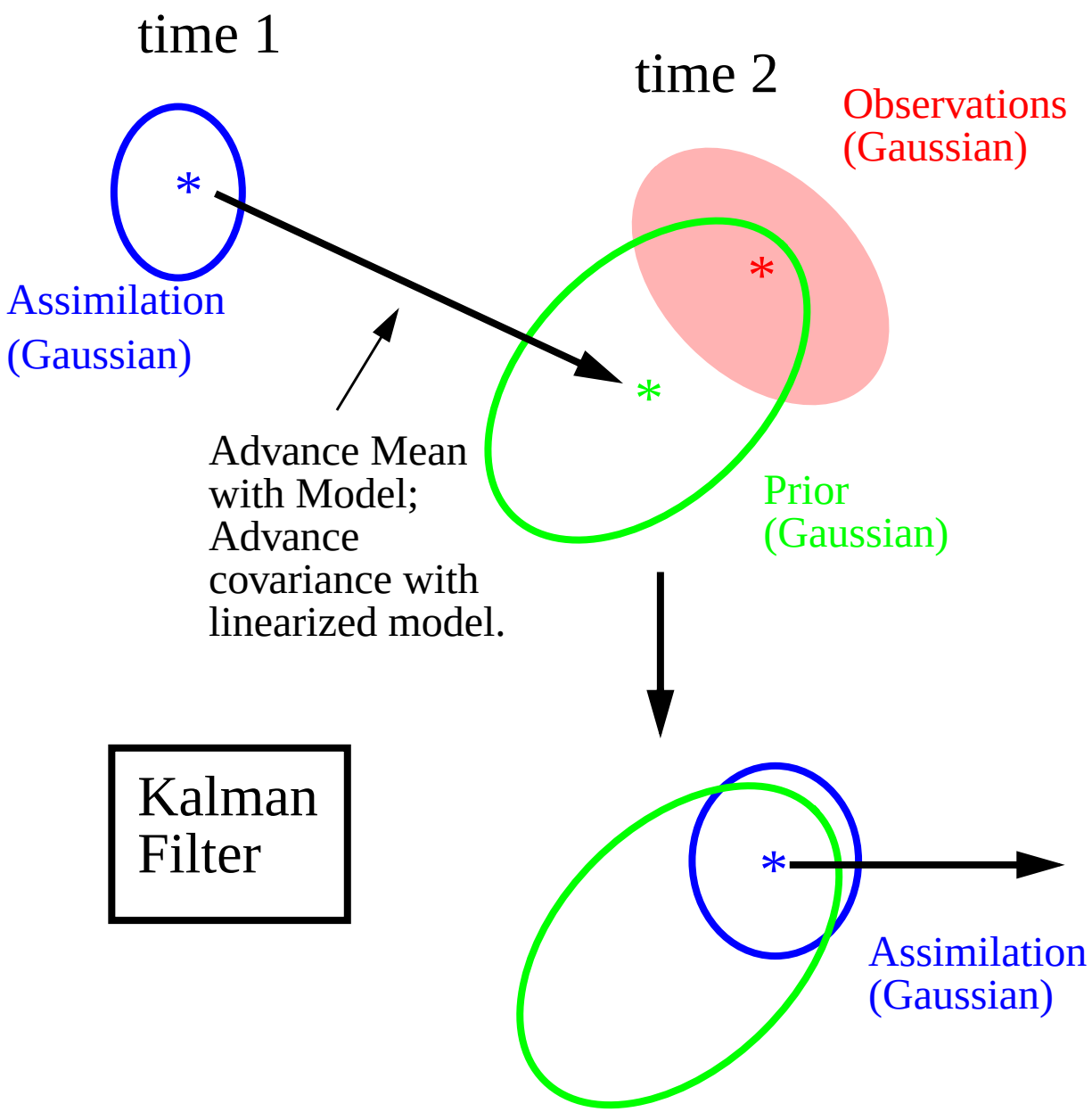
1. Linearization of model around non-linear control trajectory

2. Error distributions assumed Gaussian

Fundamental Problem:

Still too expensive for large models

Advancing covariance in linearized model is at least
O(model_size * model_size)


time 1
time 2
Observations
(Gaussian)
Assimilation
(Gaussian)
Advance Mean
with Model;
Advance
covariance with
linearized model.
Prior
(Gaussian)
Kalman
Filter
Assimilation
(Gaussian)

Reduced Space Kalman Filters:

Additional simplification:

Assume that covariance projects only on small subspace of model state

Evolving covariance in linearized model projected on subspace may be cheap

Subspace selection:

1. Dynamical: use simplified model based on some sort of scaling

2. Statistical: use long record of model (or physical system) to find reduced basis in which most variance occurs (EOF most common to date)

Problems:

1. Dynamics constrained to subspace may provide inaccurate covariance evolution

2. Observations may not ‘project strongly’ on subspace

3. Errors orthogonal to subspace unconstrained, model bias in these directions can quickly prove fatal

Ensemble Kalman Filters:

Simplifications:

1. Monte Carlo approximation to probability distributions

2. Localization in space, avoids degeneracy from samples smaller that state space

3. Gaussian representation of probability distributions generally used for computing update

Problems:

1. Selecting initial samples for ensembles (Monte Carlo samples)

2. Determining degree of spatial localization

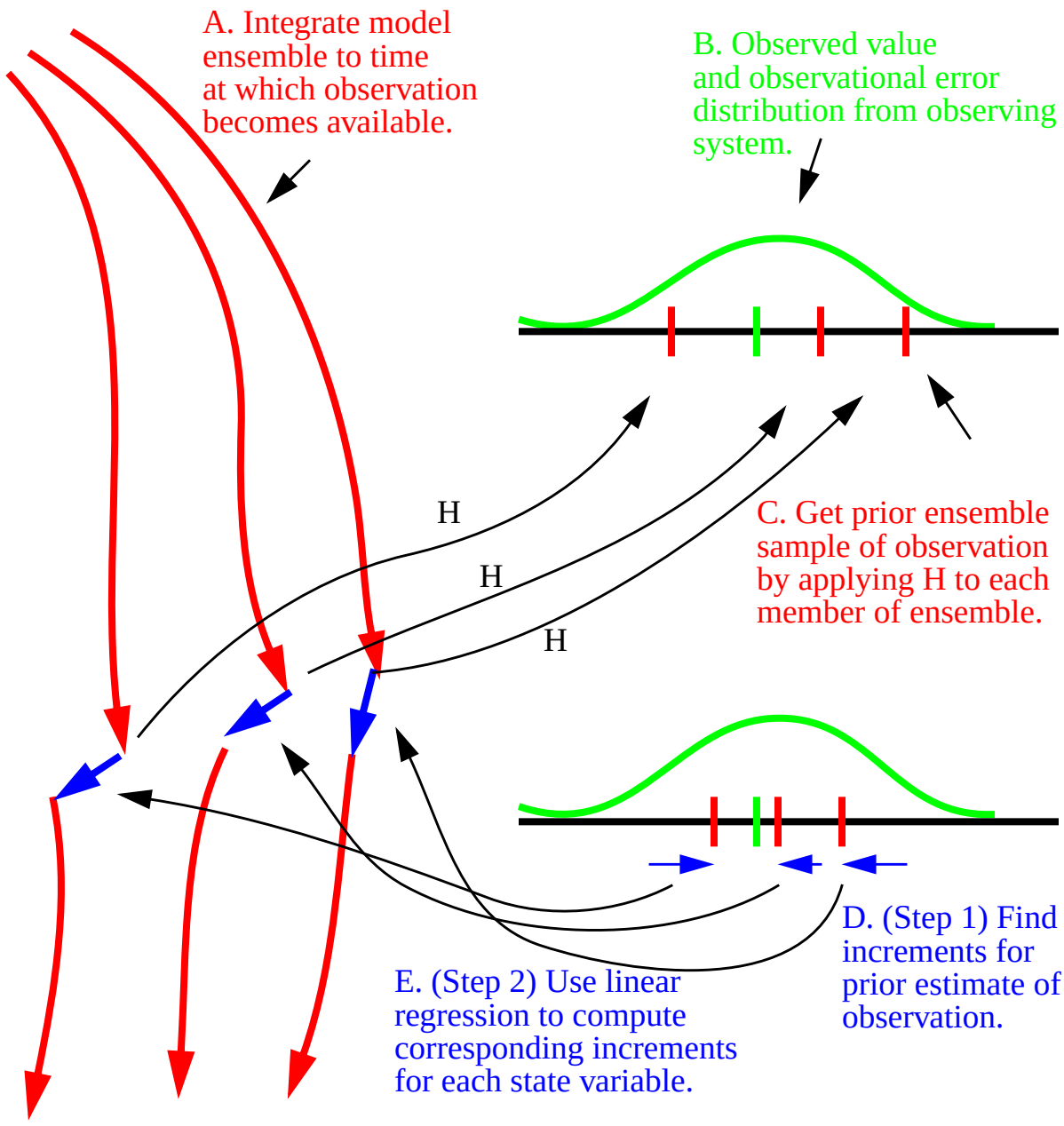
3. Maintaining appropriate model ‘balances’ in ensemble members

BUT, UNPRECEDENTED EASE OF INITIAL APPLICATION

How an Ensemble Filter Works

Theory: Impact of observations can be handled sequentially

Impact of observation on each state variable can be handled sequentially



Details of Step 1: Updating Observation Variable Ensemble

Scalar Problem: Wide variety of options available and affordable

Begin with two previously documented methods:

1. Perturbed Observation Ensemble Kalman Filter

2. Ensemble Adjustment Kalman Filter

Both make use of following (key to Kalman filter...)

Given prior ensemble with sample mean $\bar{z}^p$ and covariance Σ^p

Observation y^o with observational error variance matrix R

Note: Product of Gaussians is Gaussian

$$\Sigma^u = \left\{ \left(\Sigma^p\right)^{-1} + H^T R^{-1} H \right\}^{-1} \qquad (9)$$

and mean:

$$\bar{z^u} = \Sigma^u \left\{ \left(\Sigma^p\right)^{-1} \bar{z^p} + H^T R^{-1} y^o \right\} \qquad (10)$$

Details of Step 1: Perturbed Obs. Ensemble Kalman Filter

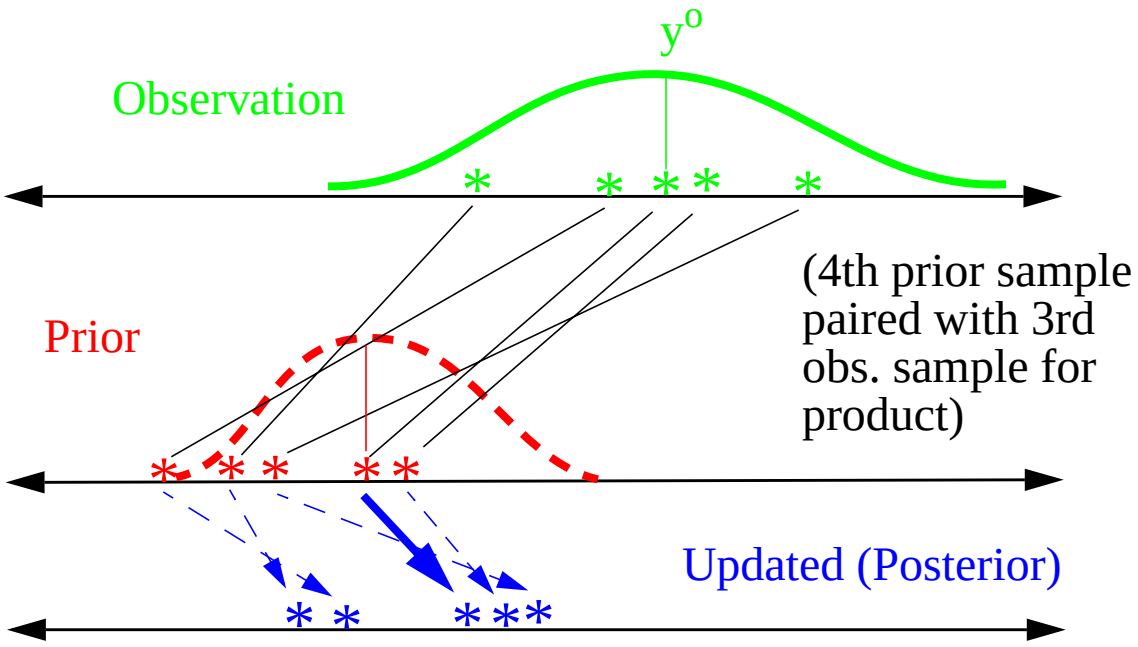
1. Compute prior sample variance and mean, Σ^p and $\bar{z}^p$

2. Apply (9) once to compute updated covariance, Σ^u

3. Create an N-member random sample of observation distribution by adding samples of obs. error to y^o

4. Apply (10) N times to compute updated ensemble members
   Replace $\bar{z}^p$ with value from prior ensemble, y^p_i
   Replace y^o with value from random sample, y^o_i
   Updated ensemble value is y^u_i (= $\bar{z}^u$ from 10)

NOTE: When combined with linear regression for step 2, this gives identical results to EnKF's described in literature!

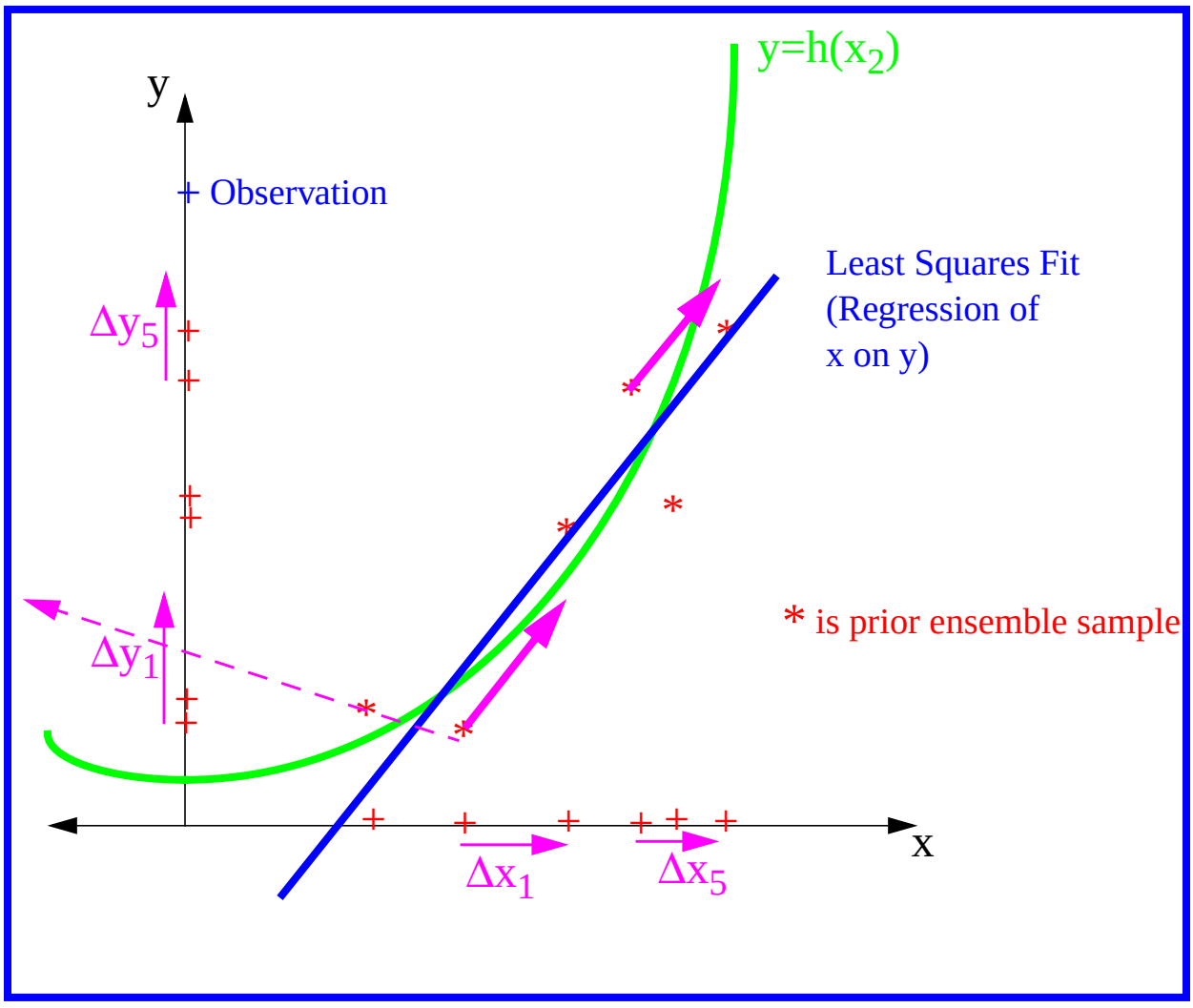


Two Step Ensemble Assimilation (cont.)

Step 2: Given increments for y, find increments for state variables

More challenging when obs and state are not functionally related

Example: $y = h(x_2)$, x and x_2 strongly correlated



Large sample size needed to 'remove' noise
Trade-offs with local linearization (dotted magenta)

Technical Difficulties Remain

Problem 1. Sampling error impacts estimates of increments

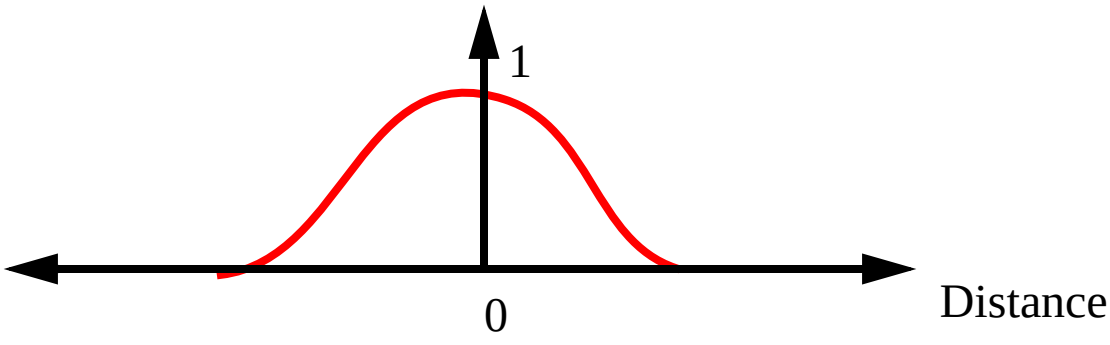
Key: estimates of correlations for regression have errors

Many obs. with small expected correlations => error build-up

Solution: Reduce impact of observations as function of ensemble size, sample correlation, and prior knowledge of expected distribution of correlation

But...need this prior estimate (may be mostly unknown?)

For now, use distance dependent envelope to reduce impact of remote observations



Even picking this envelope still tricky for now

Problem 2. Initial conditions for ensembles

Key: Bayesian, assumes initial ensembles are magically available

Solution: For ergodic models (many global GCMs) spin-up by running ensemble a very long time from arbitrary initial perturbations, slowly ‘turn on’ observations

But... this may be impossible for WRF regional applications

Given prior knowledge of expected correlations (see problem 1) should be able to generate appropriate ensemble ICs

Still a topic for ongoing research

Problem 3. Assimilation of variableswith discrete distributions

Key: ensemble prior may indicate zero probability of an event that is occurring

I.E. All ensemble members say no rain but rain is observed

Directly related to existence of discrete convective cells

Solutions: Apply methods for accounting for model error

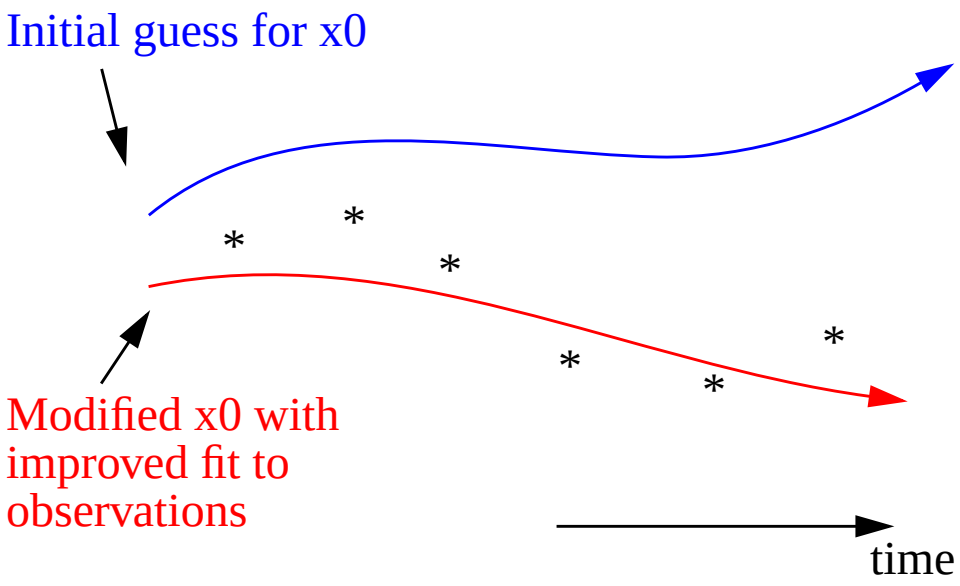
Redefine state variables to avoid discrete probability densities

Research on this problem is in its infancy

4D-Variational (4D-Var)

Find model trajectory through time that minimizes a norm measuring departure from observations

Applied over some finite period of observations



For optimization, need gradient of norm with respect to model state at initial time

Key: integrating the adjoint of the linear tangent model linearized around forward non-linear model trajectory backward in time allows computation of gradient with single integration pair

This makes 4D-Var feasible as long as period is short and number of iterations needed for optimization is small

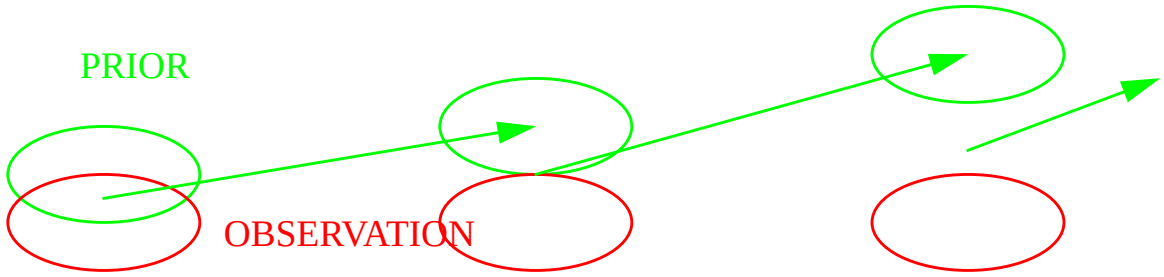
Additional problems:
1. Model 'balance' constraints may not be satisfied for finite optimization periods
2. Still hard to generate adjoints for complicated models
3. May need to relax constraints to deal with model BIAS

Challenge #1: Model Bias and Atmospheric Balances

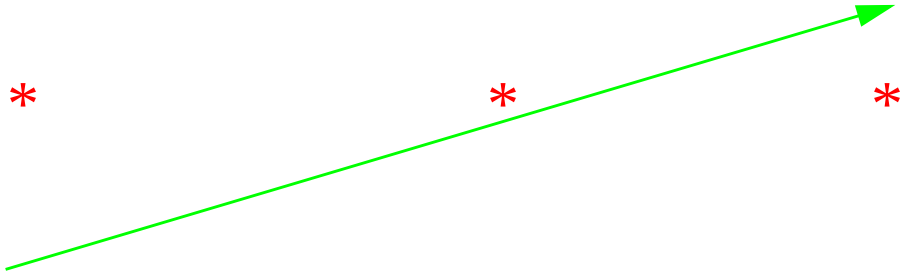
Filter equations assume prior estimate (and observations) are unbiased
Questionable for Observations, ridiculous for Models

Biased prior estimate will cause observations to be given too little weight

Repeated applications lead to progressively less weight, estimate can diverge



Implications are obvious for 4D-Var, too



Dealing with model bias is mostly an open question:

1. Can reduce confidence in model prior estimates by some constant factor

2. Explicitly model the model bias as an extended state vector and assimilate coeffi-
   cients of this bias model

Model: $dx/dt = F(x)$

Model plus bias model: $dx/dt = F(x) + \varepsilon(t)$; $d\varepsilon/dt = 0$
where ε is a vector of the same length as x

Very tricky: if we knew much about modeling the bias, we could remove

Challenge #2: Balances and Attractors

Many models of interest have balances, both obvious (say geostrophic) and subtle

The structure of the model ‘attractors’ may be extremely complex

In some cases, off-attractor perturbations may lead to ‘large’ transient response

Example: High frequency gravity waves in some Primitive Equation models

The behavior of these transients can lead to model bias

In this sense, even perfect model experiments can have large model bias

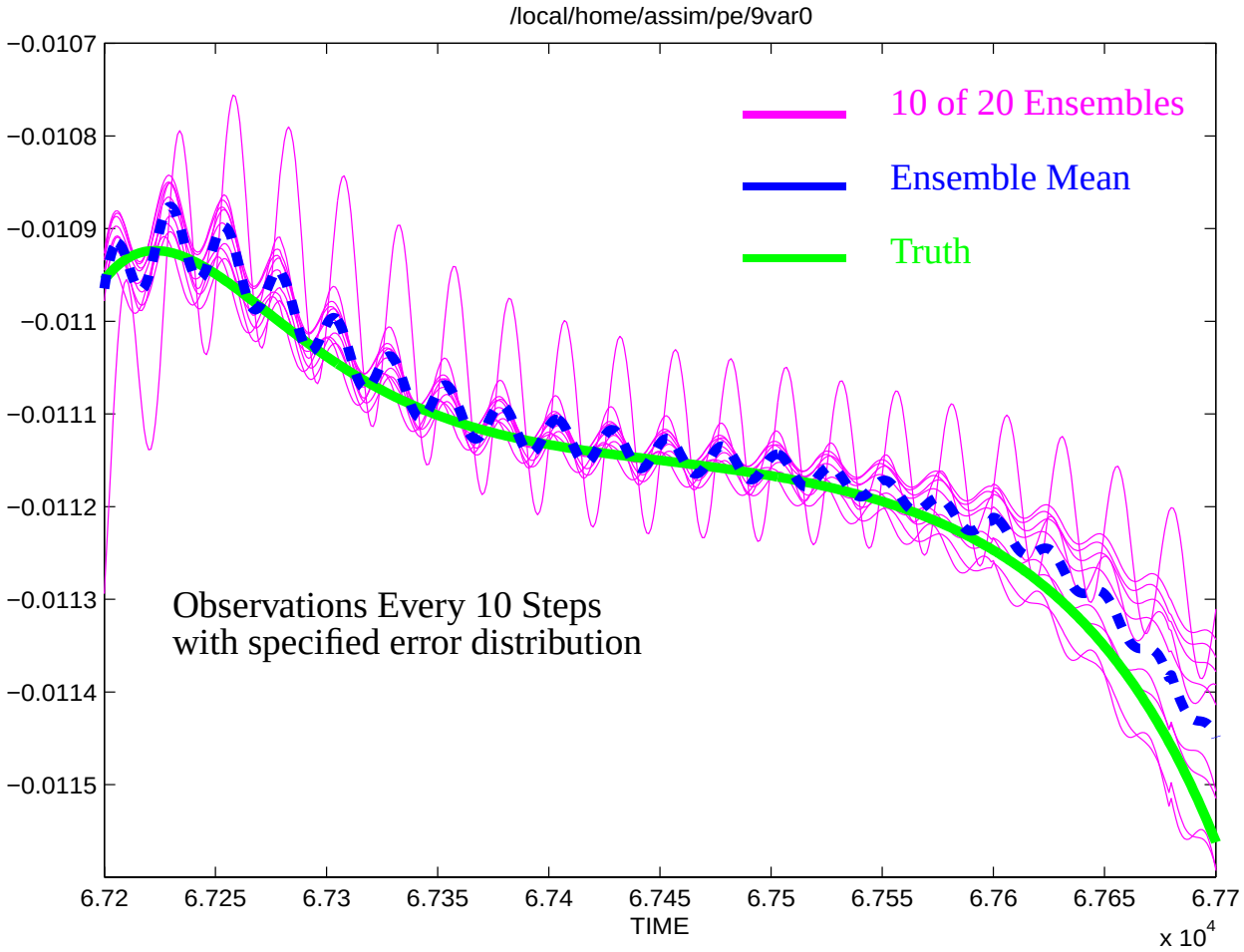
Understanding how to minimize this behavior or limit its impact is a fun problem

The continuous system may also have balances, obvious and subtle

Unclear how differences between model and continuous ‘attractors’ impacts assimilation

Lorenz 9-Variable Model

Time series of Ensemble Filter Assimilation for variable X1



$$\dot{X}_i = U_j U_k + V_j V_k - v_0 a_i X_i + Y_i + a_i z_i \quad (1)$$

$$\dot{Y}_i = U_j Y_k + Y_j V_k - X_i - v_0 a_i Y_i \quad (2)$$

$$\dot{z}_i = U_j (z_k - h_k) + (z_j - h_j) V_k - g_0 X_i - K_0 a_i z_i + F_i \quad (3)$$

$$U_i = -b_j x_i + c y_i \quad (4)$$

$$V_i = -b_k x_i - c y_i \quad (5)$$

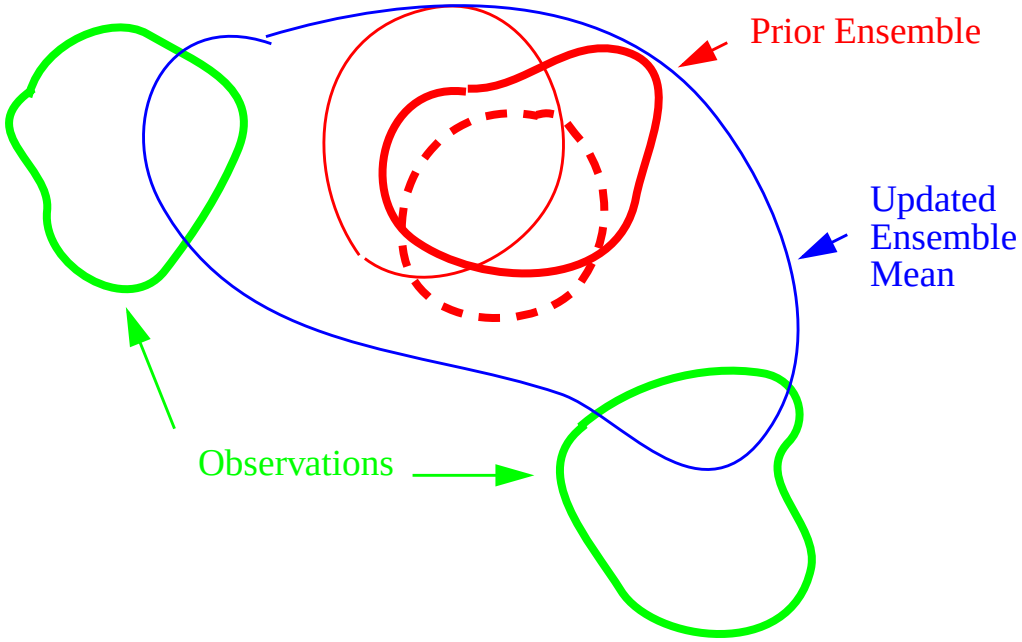
$$X_i = -a_i x_i \quad (6)$$

$$Y_i = -a_i y_i \quad (7)$$

Defined for cyclic permutations of the indices (i, j, k) over the values (1, 2, 3).
X, Y and z variables can be thought of as representing divergence, vorticity and height

Challenge #3: Assimilation of Discrete Distributions

Example: assimilation of convective elements



Prior is ‘certain’ that there are no convective cells outside the red areas
Observations indicate discrete areas outside the red
This is indicative of highly non-linear problem
Ensemble techniques, at best, tend to smear out prior discrete structures
4D-Var is likely to have non-global local minima

But, we think we know what we want to do
Keep information from prior on larger scale ‘background’
Introduce cells where observed

Requires new norms or ways to deal with model bias as function of scale

Exciting Opportunities Abound in Data Assimilation

Field is maturing, basic theory well-understood
Increasingly powerful heuristic methods being developed
Some new methods (like filter) are very simple to implement (naively)
Software engineering advances make it easier to access models and data
NCAR/NOAA are building a prototype facility for exploring DA

1. The challenges are opportunities!

2. Plethora of models and observations that have not been touched!

3. Improved assimilation application to existing high profile problems!

Example: Getting more from existing data, surface pressure observations

Example: Quality control of observations: using good data, rejecting bad

4. Using data to improve models!

Example: Application in simple low-order model

5. Stochastic (ensemble) prediction

6. Evaluating and designing observing systems! (Observing / Assimilation System Simulation Experiments)

What is information content of existing observations?

What is value of additional proposed observations?

Use of targeted (on demand) observations
Potential for extremely high impact (if you can stand the heat)

Quality Control of Observations

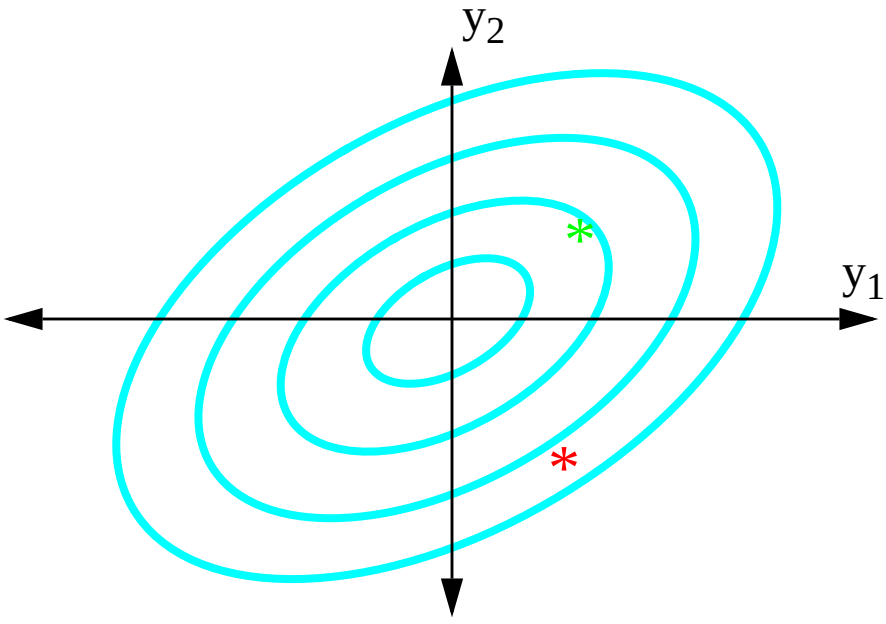
Methods to exclude erroneous observations

1. Discard impossible values (negative R.H.)

2. Discard values greatly outside climatological range

3. Discard values that are more than α prior ensemble sample standard deviations away from prior ensemble mean

4. 'Buddy' checks for pairs of observations: just apply chi-square test using prior ensemble covariance and label pair as inconsistent if threshold value exceeded



5. Could also apply chi-square to larger groups of obs.

Lorenz-96 Model

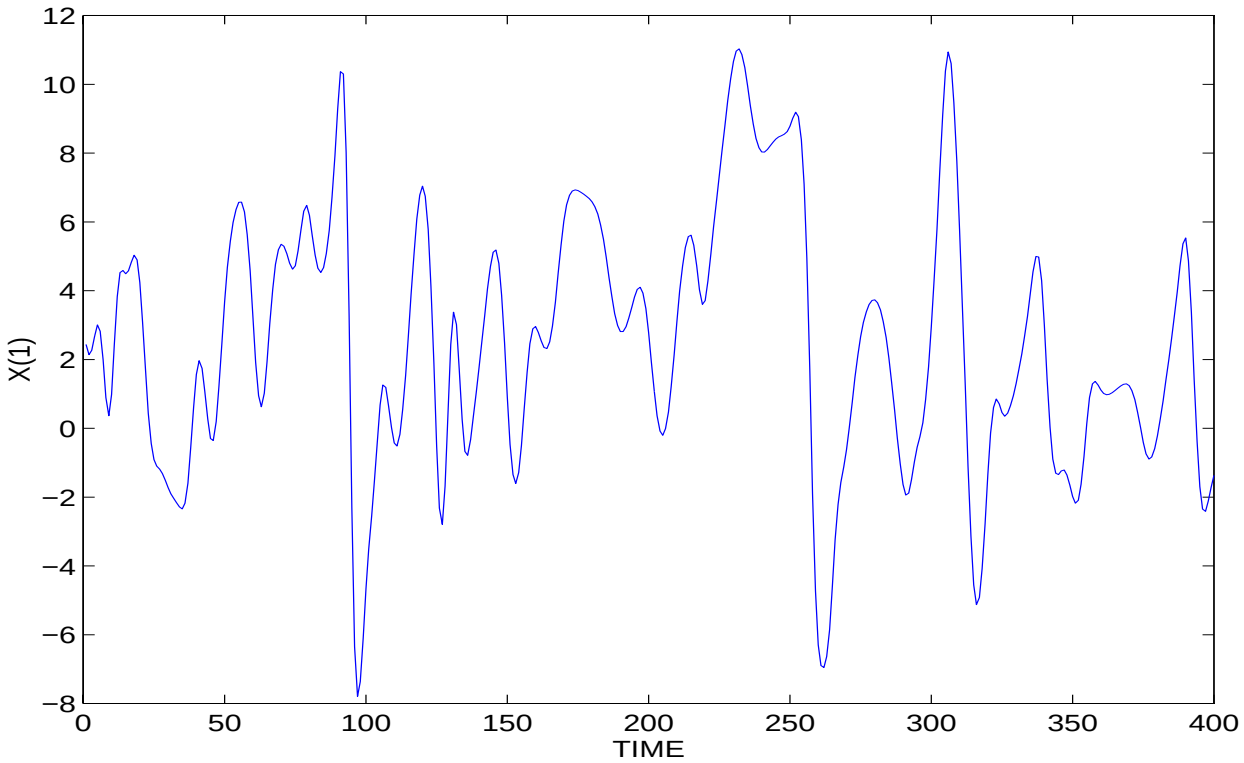
Variable size low-order dynamical system

N variables, X_1, X_2, ..., X_N

$$dX_i / dt = (X_{i+1} - X_{i-2})X_{i-1} - X_i + F$$

i = 1, ..., N with cyclic indices

Use N = 40, F = 8.0, 4th-order Runge-Kutta with dt=0.05



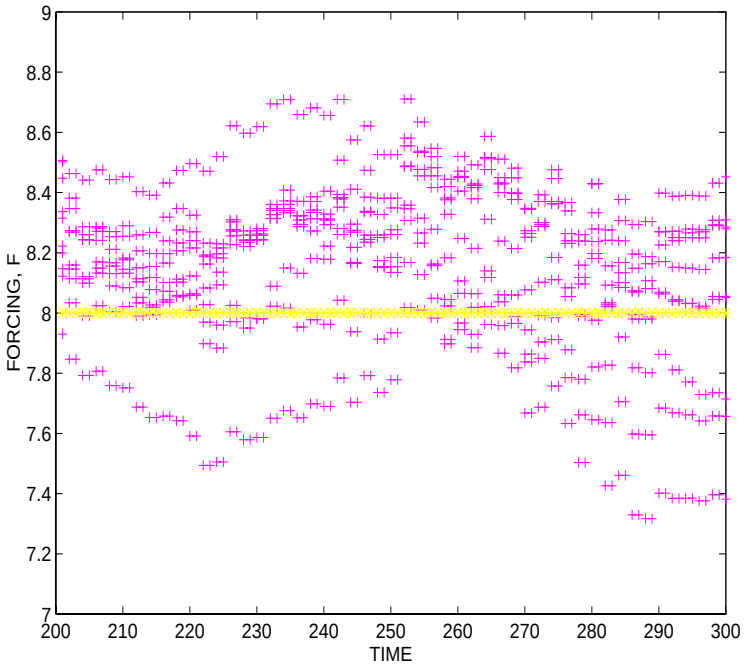
Lorenz-96 Free Forcing Model Filter

20 Member Ensemble (10 Plotted)

Obs Every 2 Steps

Truth (8.0)

Ensemble

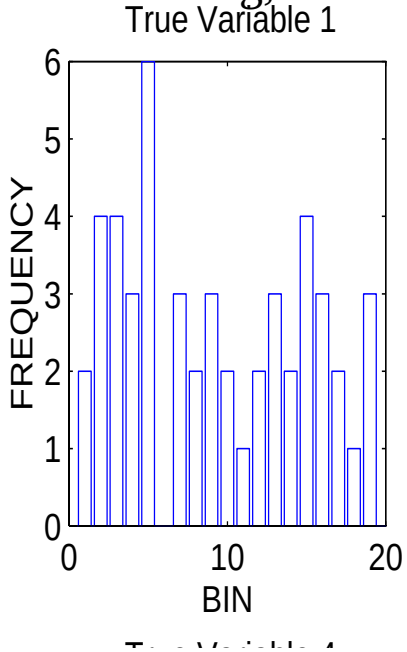


>> Can treat model parameters as free parameters <<

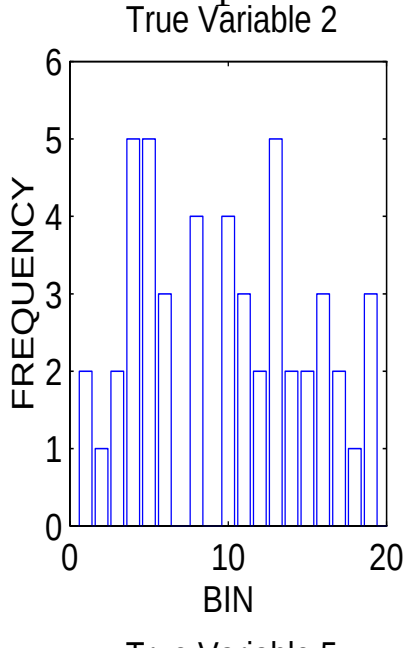
>> Here, the forcing F is assimilated along with the state <<

>> This is potential mechanism for dynamic adjustment of unknown parameters and for dealing with unknown model systematic error <<

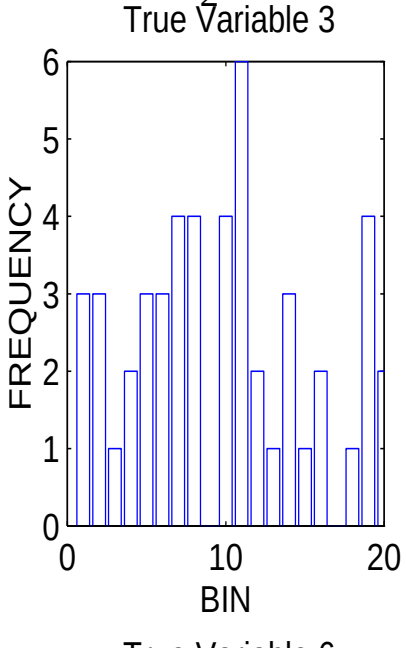
Forcing, F



X_1



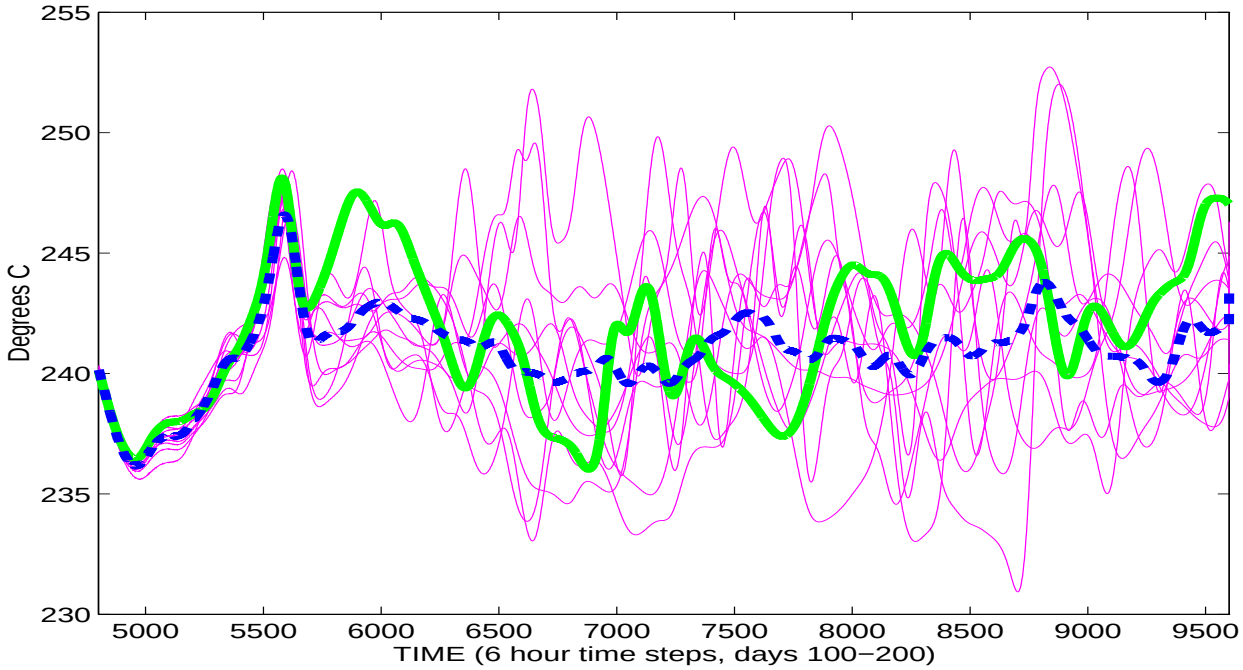
X_2



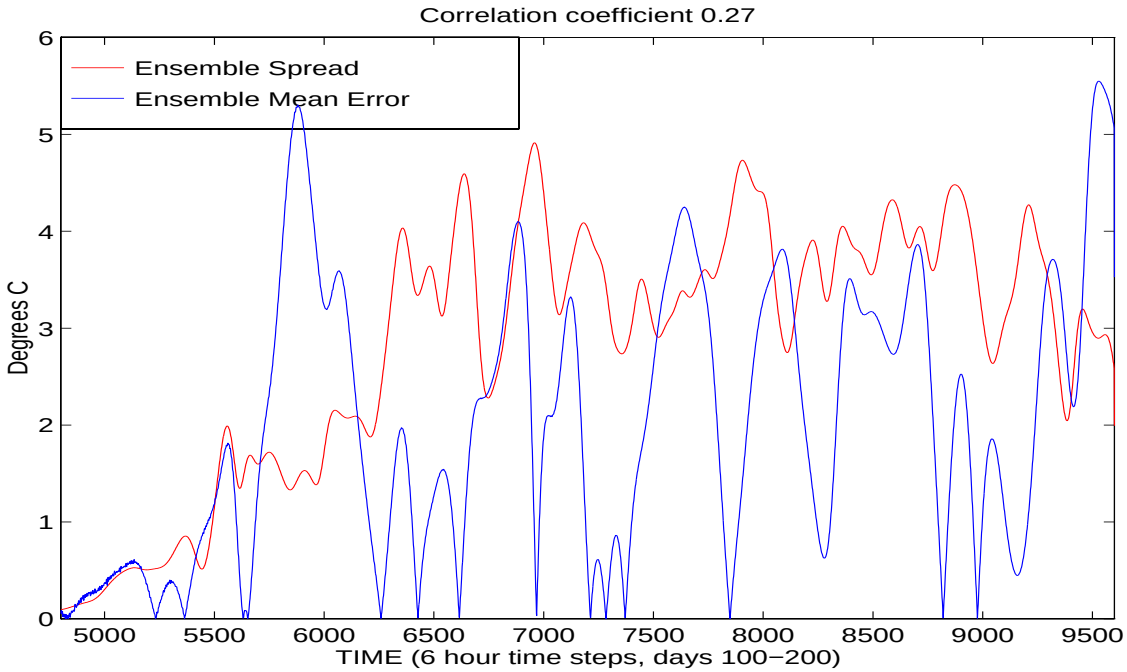
Global Dry PE Model Assimilation (40 lons x 30 lats x 5 levels)

Mid-latitude / Mid-troposphere Temperature Days 100-200

No assimilation after day 100



10 Members of 20 Member Ensemble Ensemble Mean Truth

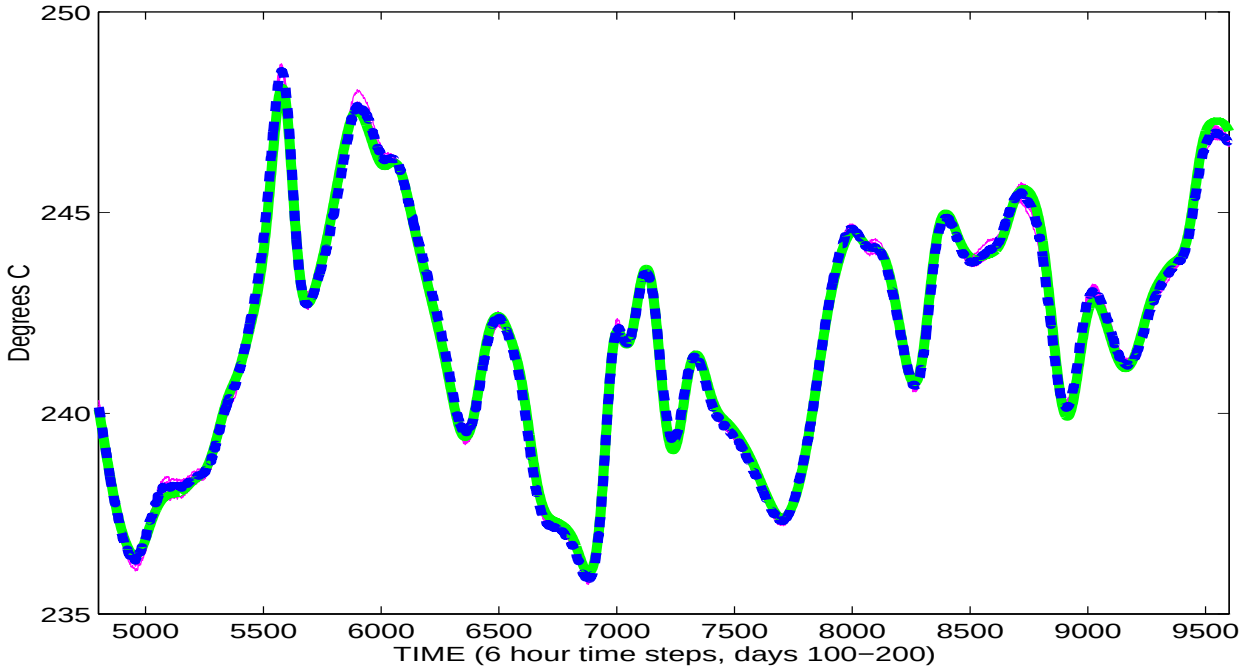


RMS Error of Ensemble Mean Ensemble Spread

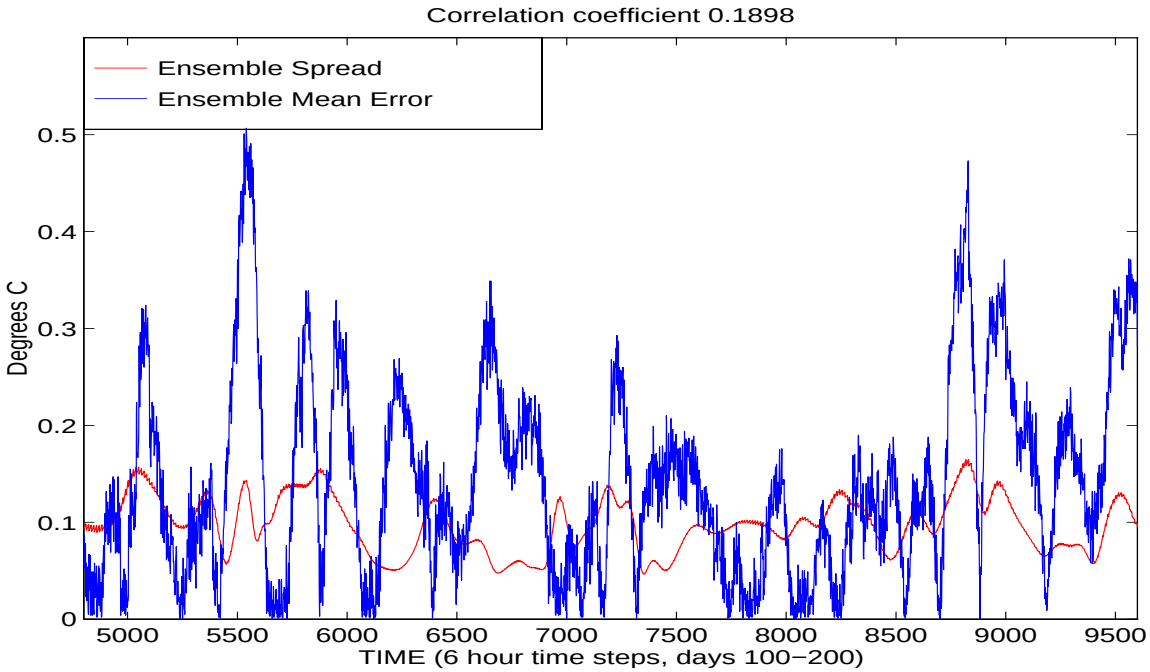
Global Dry PE Model Assimilation (40 lons x 30 lats x 5 levels)

Mid-latitude / Mid-troposphere Temperature Days 100-200

Assimilating ONLY Surface Pressure (Obs. Error S.D. 100 Pa) Every 6 Hours



10 Members of 20 Member Ensemble Ensemble Mean TruthRMS



Error of Ensemble Mean Ensemble Spread